WEB MINING UNTUK PENCARIAN DOKUMEN BAHASA INGGRIS MENGGUNAKAN HILL CLIMBING AUTOMATIC CLUSTER

Hervilorra Eldira[1], Entin Martiana K[2]., S.Kom M.Kom, Nur Rosyid M[2]., S.Kom
[1] Mahasiswa, [2] Dosen Pembimbing

Politeknik Elektronika Negeri Surabaya
Institut Teknologi Sepuluh Nopember Kampus ITS Keputih Sukolilo Surabaya 60111, Indonesia
Tel: +62-85655616926, 0351-369234
Email: hervilorra@gmail.com

Abstrak

Web mining untuk pencarian dokumen bahasa Inggris menggunakan Hill Climbing Automatic Clustering adalah sebuah aplikasi dari salah satu metode pencarian dokumen berdasarkan kata kunci yang dimasukkan oleh pemakai.

Kata kunci yang dimasukkan akan diproses menggunakan text mining untuk menghilangkan kata yang tidak berguna dan mendapatkan kata dasar. Selain itu, pada dokumen dilakukan juga text mining dan perhitungan jumlah kata, dari jumlah kata tersebut dilakukan pengklusteran dengan metode CLHM (Centroid Linkage Hierarchical Method). Untuk jumlah klusternya, pemakai tidak mengetahui berapa jumlah yang tepat untuk mengklusterkan dokumen-dokumen tersebut. Untuk itu, dipakailah metode Hill Climbing yang bertugas untuk melakukan identifikasi terhadap pergerakan varian dari tiap tahap pembentukan kluster dan menganalisa polanya sehingga jumlah kluster akan terbentuk secara otomatis.

Penggunaan text mining, pengklusteran dengan CLHM dan proses Hill Climbing Automatic Clustering sangat memudahkan pemakai karena menghasilkan kluster secara otomatis dan tepat dengan waktu yang cepat.

Kata kunci : *Text Mining, Automatic Clustering, Centroid Linkage Hierarchical Method, Hill Climbing, Online Clustering*

1. PENDAHULUAN

1.1. LATAR BELAKANG

Penyimpanan dokumen secara *digital* berkembang dengan pesat seiring meningkatnya penggunaan komputer. Kondisi tersebut memunculkan masalah untuk mengakses informasi yang diinginkan secara akurat dan cepat. Oleh karena itu, walaupun sebagian besar dokumen *digital* tersimpan dalam bentuk teks dan berbagai algoritma yang efisien untuk pencarian teks telah dikembangkan, teknik pencarian terhadap seluruh isi dokumen yang tersimpan bukanlah solusi yang tepat mengingat pertumbuhan ukuran data yang tersimpan umumnya.

Pencarian informasi (*Information Retrieval*) [1] adalah salah satu cabang ilmu yang menangani masalah ini yang bertujuan untuk membantu pengguna dalam menemukan informasi yang relevan dengan kebutuhan

mereka dalam waktu singkat. Aplikasi pencarian informasi yang telah ada salah satunya adalah *web mining* untuk pencarian berdasarkan kata kunci dengan teknik *clustering*.

Dalam proyek akhir sebelumnya telah dibuat suatu aplikasi *web mining* untuk pencarian berdasarkan kata kunci dengan *outomatic clustering* untuk mengelompokkan dokumen bahasa Indonesia. Dengan menggunakan teknik pengklasteran berdasarkan algoritma *Centroid Linkage Hierarchical Method* dan analisa pola *varian* yang memenuhi *valley tracing* maka dokumen dapat diklasterkan dengan jumlah *cluster* yang tepat secara otomatis.

Pada proyek akhir yang lalu dokumen yang dijadikan sebagai sumber adalah dokumen berbahasa Indonesia dan masih *offline*. Maka pada proyek akhir ini akan dibuat suatu rancangan aplikasi *web mining* untuk pencarian berdasarkan kata kunci dengan *outomatic clustering* untuk mengelompokkan dokumen bahasa Inggris dengan menggunakan teknik pengklasteran berdasarkan algoritma *Centroid Linkage Hierarchical Method* dan analisa pola *varian* yang memenuhi *Hill-climbing* dan dalam implementasinya akan dibuat secara online.

1.2. PERMASALAHAN

Pada proyek akhir untuk membangun sistem mesin pencari ini akan dibahas permasalahan yang penting yaitu:

- Bagaimana analisa pola *varian* yang memenuhi *Hill-climbing* dapat membentuk jumlah *cluster* dokumen secara tepat dan otomatis.
- Bagaimana mengimplementasikan fungsi *text mining* yang dioptimalkan sebagai sarana untuk pencarian dokumen bahasa Inggris.
- Bagaimana mengimplementasikan *clustering* dokumen dengan cara *online*.

1.3. BATASAN MASALAH

Dalam proyek akhir ini permasalahan difokuskan pada masalah-masalah berikut :

- Penggunaan dokumen yang berbahasa Inggris sebagai dokumen yang akan diolah.
- Penggunaan aplikasi secara *online* dalam pencarian dan penyimpan dokumen yang akan di*cluster*.
- Dokumen yang akan diolah oleh *cluster* diambil dari beberapa situs berita *online*.
- Dilakukan proses normalisasi pada data yang akan di*cluster* yang mempunyai perbedaan *range* yang besar.

1.4. TUJUAN DAN MANFAAT

Proyek akhir yang berjudul "*Web Mining untuk Pencarian Dokumen Bahasa Inggris Menggunakan Hill Climbing Automatic Clustering*" ini dirancang bertujuan untuk membuat sebuah aplikasi *clustering engine* (mesin pengkluster) yang dapat mengelompokkan dokumen dengan jumlah yang tepat secara otomatis dan menampilkan hasil pencarian dokumen yang relevan dengan kata kunci yang dicari oleh *user*.

2. TEORI PENUNJANG

2.1. TEXT MINING

Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks (*text categorization*) dan pengelompokan teks (*text clustering*).

Algoritma yang digunakan pada *text mining*, biasanya tidak hanya melakukan perhitungan hanya pada dokumen, tetapi

pada juga *feature*. Empat macam *feature* yang sering digunakan:

- *Character*, merupakan komponan individual, bisa huruf, angka, karakter spesial dan spasi, merupakan *block* pembangun pada level paling tinggi pembentuk semantik *feature*, seperti kata, *term* dan *concept*.Pada umumnya, representasi *character-based* ini jarang digunakan pada beberapa teknik pemrosesan teks.
- *Words*.
- *Terms* merupakan *single word* dan frasa *multiword* yang terpilih secara langsung dari corpus. Representasi *term-based* dari dokumen tersusun dari subset *term* dalam dokumen.
- *Concept*, merupakan *feature* yang di-generate dari sebuah dokumen secara manual, *rule-based*, atau metodologi lain. Pada tugas akhir ini, konsep di-generate dari *argument* atau *verb* yang sudah diberi label pada suatu dokumen.

Proses text mining meliputi proses *tokenizing*, *filtering*, *stemming*,dan *tagging*.

2.1.1. Tokenizing

Tokenizing adalah proses penghilangan tanda baca pada kalimat yang ada dalam dokumen sehingga menghasilkan kata-kata yang berdiri sendiri-sendiri [2].

2.1.2. Filtering

Tahap *filtering* adalah tahap pengambilan kata-kata yang penting dari hasil *tokenizing*. Tahap *filtering* ini dapat menggunakan algoritma *stoplist* atau *wordlist*. *Stoplist* yaitu penyaringan (*filtering*) terhadap kata-kata yang tidak layak untuk dijadikan sebagai pembeda atau sebagai kata kunci dalam pencarian dokumen sehingga kata-kata tersebut dapat dihilangkan dari dokumen. Sedangkan *wordlist* adalah daftar kata-kata yang mungkin digunakan sebagai kata kunci dalam pencarian dokumen, dengan demikian maka tentu jumlah kata yang termasuk dalam *wordlist* akan lebih banyak daripada *stoplist*, sehingga dalam proyek akhir ini digunakan daftar *stoplist*. Oleh karena belum tersedia, maka pada proyek akhir ini juga akan berusaha mencari *stoplist* tersebut secara manual [2].

2.1.3. Stemming

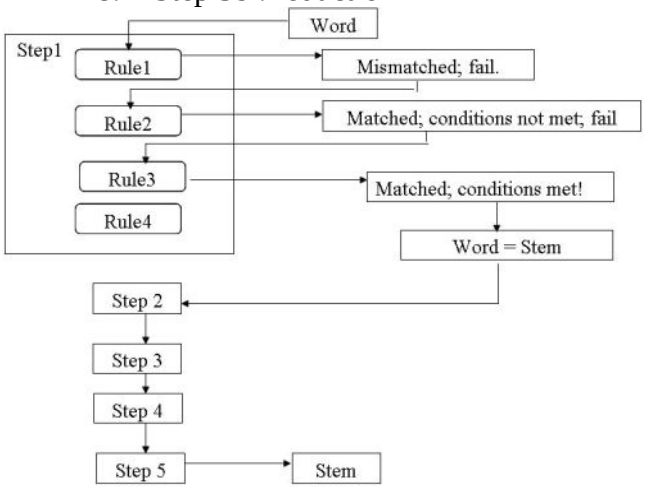
Stemming adalah proses mengubah kata menjadi kata dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen atau mengubah kata kerja menjadi kata benda [2]. Stem (akar kata) adalah bagian dari kata yang tersisa setelah dihilangkan imbuhannya (awalan dan akhiran). Contoh : *connect* adalah stem dari *connected, connecting, connection*, dan *connections*.

Gambar 1. Bagan metode *stemming*

Porter stemmer[3] merupakan algoritma penghilangan akhiran *morphological* dan *infleksional* yang umum dari bahasa Inggris. Step-step algoritma *Porter Stemmer* :

1. Step 1a : *remove plural suffixation*
2. Step 1b : *remove verbal inflection*
3. Step 1b1 : *continued for -ed and -ing rules*
4. Step 1c : y and i
5. Step 3
6. Step 4 : delete last suffix

7. Step 5a : *remove e*
8. Step 5b : reduction



Gambar 2. *Control flow* algoritma *Porter Stemmer*

2.1.4. Tagging

Tahap *tagging* adalah tahap mencari bentuk awal / root dari tiap kata lampau atau kata hasil *stemming*. Contoh : *was* → *be, used* → *use, stori* → *story,* dll.

2.2. AUTOMATIC CLUSTERING

Clustering adalah proses membuat pengelompokan sehingga semua anggota dari setiap partisi mempunyai persamaan berdasarkan matrik tertentu. Sebuah *cluster* adalah sekumpulan objek yang digabung bersama karena persamaan atau kedekatannya [4]. *Clustering* atau klasterisasi merupakan sebuah teknik yang sangat berguna karena akan mentranslasi ukuran persamaan yang intuitif menjadi ukuran yang kuantitatif.

2.2.1. CLHM(Centroid Linkage Hierarchical Method)

Centroid Linkage adalah proses pengklasteran yang didasarkan pada jarak antar *centroid*nya [6]. Metode ini baik untuk kasus *clustering* dengan normal data set *distribution*. Akan tetapi metode ini tidak cocok untuk data yang mengandung *outlier*. Algoritma *Centroid Linkage Hierarchical Method* adalah sebagai berikut:

1. Diasumsikan setiap data dianggap sebagai *cluster*. Kalau *n*=jumlah data dan *c*=jumlah *cluster*, berarti ada *c*=*n*.
2. Menghitung jarak antar *cluster* dengan *Euclidian distance*.
3. Mencari 2 *cluster* yang mempunyai jarak *centroid* antar *cluster* yang paling minimal dan digabungkan (*merge*) kedalam *cluster* baru (sehingga *c*=*c*-1).
4. Kembali ke langkah 3, dan diulangi sampai dicapai *cluster* yang diinginkan.

2.2.2. Analisa Cluster

Analisa *cluster* adalah suatu teknik analisa *multivariate* (banyak variabel) untuk mencari dan mengorganisir informasi tentang variabel tersebut sehingga secara relatif dapat dikelompokkan dalam bentuk yang homogen dalam sebuah *cluster* [5]. Secara umum, bisa dikatakan sebagai proses menganalisa baik tidaknya suatu proses pembentukan *cluster*. Analisa *cluster* bisa diperoleh dari kepadatan *cluster* yang dibentuk (*cluster density*). Kepadatan suatu *cluster* bisa ditentukan dengan *variance within cluster* (Vw) dan *variance between cluster* (Vb).

Varian tiap tahap pembentukan *cluster* bisa dihitung dengan rumus:

$$Vc^2 = \frac{1}{n_c - 1}\sum_{i=1}^{n}\left(y_i - \bar{y}_c\right)^2 \quad \ldots(1)$$

Dimana:
Vc^2 = varian pada *cluster c*
c = 1..k, dimana k = jumlah cluster
n_c = jumlah data pada *cluster c*
y_i = data ke-*i* pada suatu cluster
$\overline{y_i}$ = rata-rata dari data pada suatu *cluster*

Selanjutnya dari nilai varian diatas, kita bisa menghitung nilai *variance within cluster* (Vw) dengan rumus:

$$Vw = \frac{1}{N-c}\sum_{i=1}^{c}(n_i - 1)\cdot Vi^2 \quad \ldots(2)$$

Dimana, *N* = Jumlah semua data
n_i = Jumlah data *cluster* i
Vi= Varian pada *cluster* i

Dan nilai *variance between cluster* (Vb) dengan rumus:

$$Vb = \frac{1}{c-1}\sum_{i=1}^{c} n_i\left(\overline{y}_i - \overline{y}\right)^2 \quad (3)$$

Dimana, $\overline{y}$ = rata-rata dari $\overline{y_i}$

Salah satu metode yang digunakan untuk menentukan *cluster* yang ideal adalah batasan *variance*, yaitu dengan menghitung kepadatan *cluster* berupa *variance within cluster* (V_w) dan *variance between cluster* (V_b) [5]. *Cluster* yang ideal mempunyai V_w minimum yang merepresentasikan *internal homogenity* dan maksimum V_b yang menyatakan *external homogenity*.

$$V = \frac{Vw}{Vb} \quad \ldots(4)$$

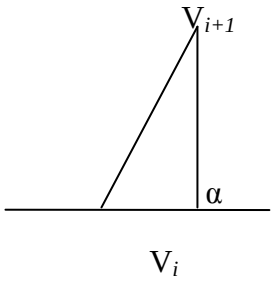
2.2.3. Hill Climbing

Pada *Hill-climbing* didefinisikan bahwa kemungkinan mencapai *global optimum* terletak pada tahap ke-*i*, jika memenuhi persamaan berikut:

$$V_{i+1} > \alpha . V_i \quad \ldots\ldots\ldots (5)$$

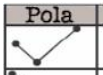
Dimana, α adalah nilai tinggi.
Nilai tinggi digunakan untuk menentukan seberapa mungkin metode ini mencapai *global optimum*. Nilai α yang biasa digunakan adalah 2,3, dan 4.

Persamaan diatas, diperoleh berdasar analisa pergerakan varian pola *Hill-climbing* yang ditunjukkan pada gambar 3 berikut:



Gambar 3. Pola nilai beda *Hill-climbing*

Berikut tabel 1 yang menunjukkan pola-pola *valley tracing* dan *hill climbing* yang mungkin mencapai global optimum [6]. Pola yang mungkin ditandai dengan simbol √.

Pola	Mungkin	Pola	Mungkin
	√		X
	√		X
	√		X
	X		√
	X		X
	X		X
	X		X
	X		X

Tabel 1. Tabel kemungkinan pola *hill climbing* mencapai global optimum

Selanjutnya, dengan pendekatan metode *hill climbing* dilakukan identifikasi perbedaan nilai tinggi (∂) pada tiap tahap, yang didefinisikan dengan :

$$\partial = V_{i+1} - (V_i * \alpha) \qquad \ldots\ldots(6)$$

Nilai ∂ digunakan untuk menghindari *local optima*, dimana persamaan ini diperoleh dari maksimum ∂ yang dipenuhi pada persamaan 6.
Untuk membentuk *cluster* secara otomatis, yaitu *cluster* yang mencapai *global optima*, digunakan nilai λ sebagai *threshold*, sehingga *cluster* secara otomatis terbentuk ketika memenuhi :

$$\max(\partial) \geq \lambda \qquad \ldots\ldots\ldots\ldots\ldots(7)$$

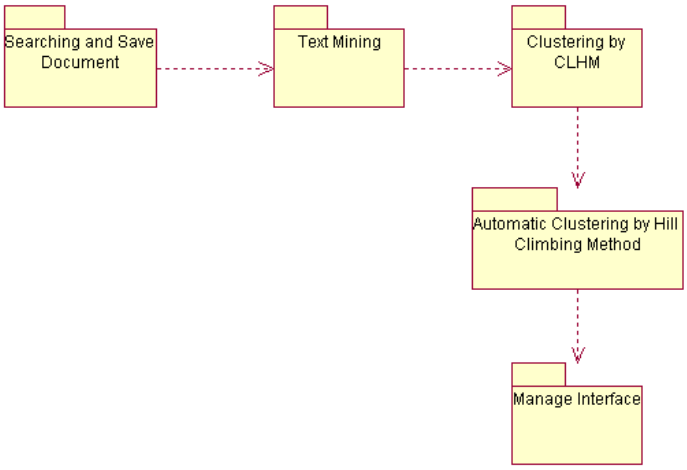
Untuk mengetahui keakuratan dari suatu metode pembentukan *cluster* pada *hierarchical method*, dengan menggunakan *hill climbing* digunakan persamaan sebagai berikut :

$$\varphi = \frac{\max(\partial)}{\text{Nilai terdekat ke max}(\partial)} \qquad (8)$$

Dimana *nilai terdekat ke max*(∂) adalah nilai kandidat max(∂) sebelumnya. Nilai φ yang lebih besar atau sama dengan 2 ($\varphi \geq 2$), menunjukkan *cluster* yang terbentuk merupakan *cluster* yang *well-separated* (terpisah dengan baik).

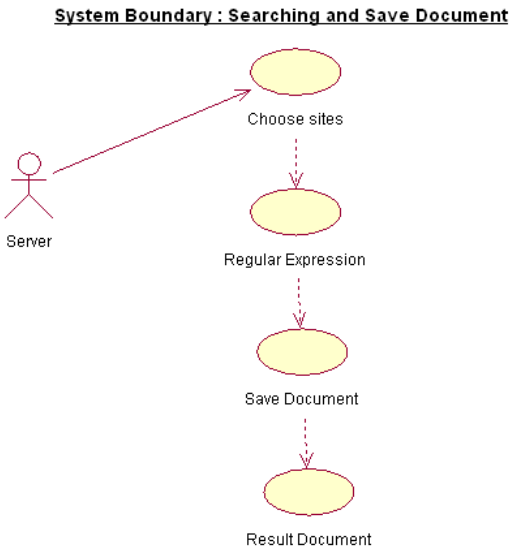
3. PERANCANGAN DAN IMPLEMENTASI

Use-Case Utama (*Architecturally Significant*) pada gambar 4 adalah gambaran sistem secara garis besar yang dibedakan menjadi lima proses utama, yaitu proses *searching* dan simpan dokumen online, proses *text mining*, proses pengklasteran dengan algoritma *Centroid Linkage Hierarchical Method*, proses pembentukan jumlah *cluster* secara otomatis (*automatic clustering*) dan bagaimana menampilkan hasil pencarian dokumen.



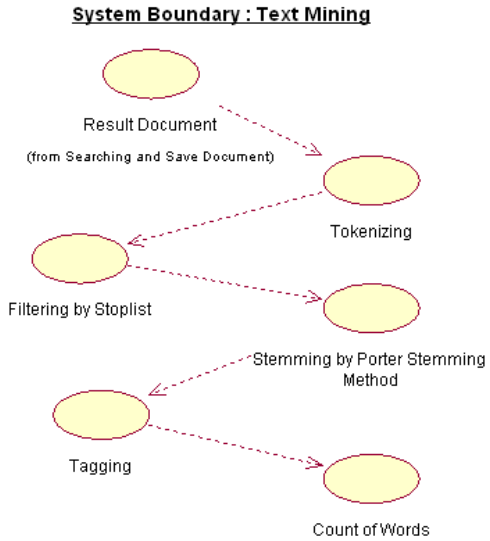
Gambar 4. Use case utama *(Architecturally Significant)*

Gambar 5 adalah merupakan *use-case diagram* untuk proses pencarian dan penyimpanan dokumen yang diambil dari internet.



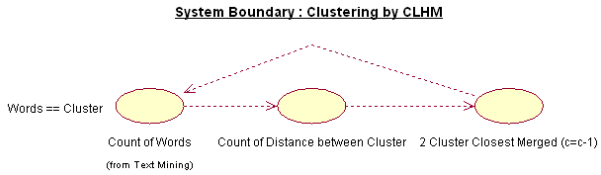
Gambar 5 Use case diagram proses pencarian dan penyimpanan dokumen dari internet

Gambar 6 adalah *use-case diagram* untuk proses text mining dimana *user* yang akan melakukan pencarian dokumen harus memasukkan *keywords* (kata kunci) terlebih dahulu kemudian sistem akan melakukan proses dari text mining.



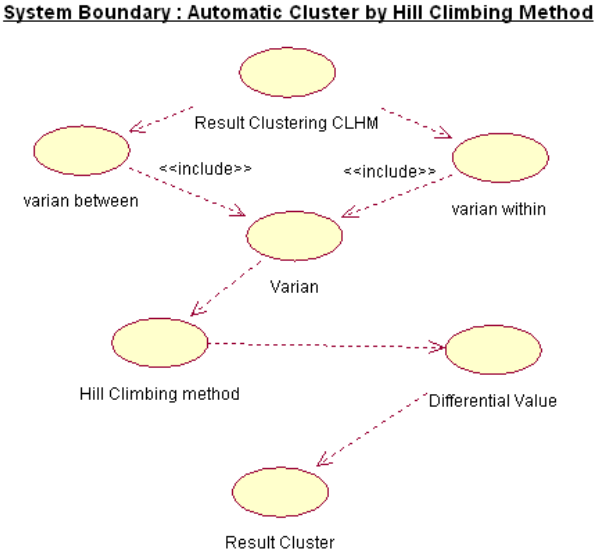
Gambar 6. Use case diagram proses *text mining*

Gambar 7 menunjukkan proses *clustering* dengan menggunakan metode CLHM (*Centroid Linkage Hierarchical Method*). Kata kunci yang dimasukkan oleh user akan dicari jumlahnya oleh sistem pada dokumen kemudian jumlah ini yang akan menentukan proses clustering berikut.



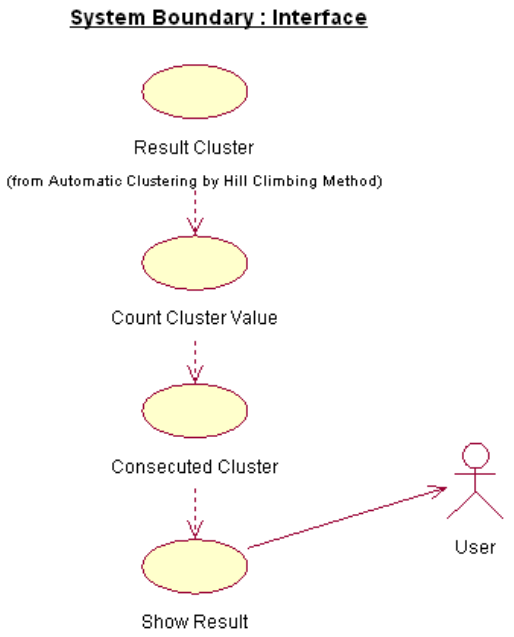
Gambar 7 Use case diagram proses *clustering* dengan CLHM

Gambar 8 menunjukkan proses dari pembentukan *automatic clustering* dengan melihat pola pergerakan varian yang ada. Dengan menggunakan metode *hill climbing* maka dianalisa posisi *global optimum* yang mungkin sehingga bisa dibentuk jumlah *cluster* yang tepat.



Gambar 8. Use case diagram proses *automatic clustering* dengan *hill climbing*

Gambar 9 menunjukkan hasil akhir dari proses *clustering engine* ini. Yaitu menampilkan hasil dokumen yang tepat sesuai dengan kata kunci yang di*input*kan oleh *user*.



Gambar 9. Use case diagram proses hasil pencarian dokumen sesuai kata kunci

4. UJI COBA DAN ANALISA

Aplikasi *Web Mining* untuk Pencarian Dokumen Bahasa Inggris Menggunakan *Hill Climbing Automatic Clustering* ini diujicobakan untuk 5 kata kunci dengan jumlah data 1000 dokumen. Hasilnya akan dibandingkan antara metode *Valley Tracing* dengan *Hill Climbing* pada proses *Automatic Clustering*nya. Dan hasilnya adalah sebagai berikut:

Keyword	jumlah cluster/masuk cluster ke-			
	valley tracing	hill climbing(α=2)	hill climbing(α=3)	hill climbing(α=4)
"Russia and US"	3/ cluster ke-3	14/cluster ke-12	14/cluster ke-12	14/cluster ke-12
"world cup"	3/cluster ke-3	97/cluster ke-78	97/cluster ke-78	97/cluster ke-78
"gaza palestine"	23/cluster ke-21	23/cluster ke-21	45/cluster ke-41	45/cluster ke-41
"Taliban Afghan"	32/cluster ke-30	41/cluster ke-39	41/cluster ke-39	41/cluster ke-39
"US President Barack Obama"	39/cluster ke-36	74/cluster ke-69	74/cluster ke-69	94/cluster ke-88

Tabel 2. Tabel perbandingan jumlah *cluster*

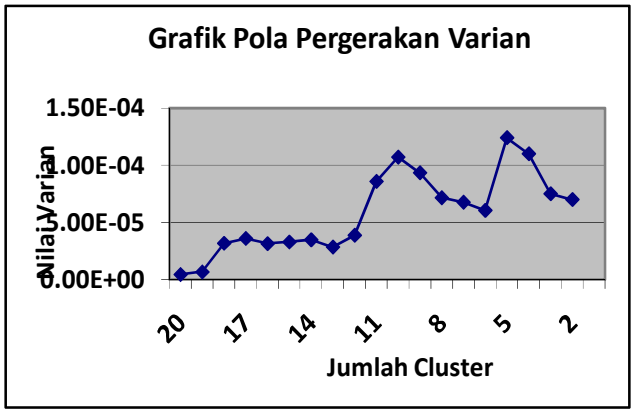
Keyword	waktu (dalam second)			
	valley tracing	hill climbing(α=2)	hill climbing(α=3)	hill climbing(α=4)
"Russia and US"	65	62	62	67
"world cup"	89	94	98	105
"gaza palestine"	80	75	84	83
"Taliban Afghan"	80	84	88	86
"US President Barack Obama"	133	125	120	105

Tabel 3. Tabel perbandingan waktu *running*

Analisa percobaan

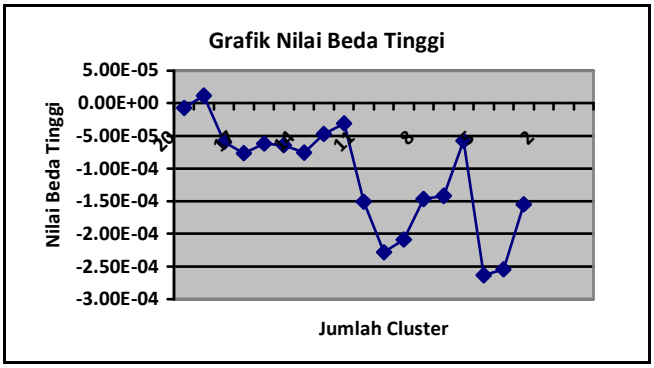
- Metode *hill climbing* dengan *threshold* 2,3,dan 4 relatif akan menghasilkan jumlah *cluster* yang sama.
- Perbedaan antara metode *hill climbing* dengan *valley tracing* sangat jelas dan jauh berbeda dalam hal jumlah *cluster* maupun waktu *running*. Untuk jumlah *cluster hill climbing* relatif menghasilkan jumlah *cluster* yang lebih banyak karena sensitivitas tiap *cluster*. Sedangkan kecepatan running disebabkan karena jumlah anggota tiap cluster sedikit sehingga proses pencarian dokumen lebih cepat.
- Untuk metode valley tracing, dalam hal waktu *running* berbanding lurus dengan jumlah kata yang diinputkan. Semakin sedikit kata kunci yang dimasukkan oleh *user*, maka waktu yang diperlukan untuk proses pengklusteran relatif cepat. Sebaliknya untuk kata kunci yang panjang lebih membutuhkan waktu yang lama. Sedangkan metode *hill climbing*, semakin banyak kata kunci yang dimasukkan, maka prosesnya relatif lebih cepat daripada metode *valley tracing*. Maka bisa disimpulkan bahwa untuk kata kunci yang panjang dan data yang banyak, sebaiknya menggunakan metode *hill climbing* untuk proses *automatic clustering.* Sedangkan untuk kata kunci yang pendek dan data yang sedikit, sebaiknya menggunakan metode *valley tracing.*

Grafik pada gambar 10 menunjukkan pola pergerakan nilai *varian* dari *cluster*. Grafik tersebut menunjukkan bahwa pola pergerakan tersebut memiliki kemungkinan untuk mencapai global optimum.



Gambar 10. Grafik Pola Pergerakan Varian

Jika dihitung dengan menggunakan rumus pada metode *hill climbing*, maka grafik nilai beda tinggi (∂) ditunjukkan pada gambar 11.



Gambar 11. Grafik Nilai Beda Tinggi

Dari grafik pada gambar 11 bisa diketahui bahwa nilai tertinggi untuk grafik ditunjukkan pada saat jumlah cluster sebanyak 19. Karena jumlah cluster yang dianggap paling optimal adalah pada saat nilai tinggi (∂) menunjukkan angka terbesar, yaitu pada saat cluster berjumlah 19, maka ini sesuai dengan metode *Hill Climbing Automatic Clustering*.

5. PENUTUP

5.1. KESIMPULAN

Dari hasil uji coba dan analisa yang telah dilakukan, maka dapat diambil kesimpulan:

1. Penggunaan *text mining* untuk pengkategorisasian teks dokumen bahasa Inggris memudahkan dalam pencarian dokumen yang sesuai dengan keinginan dari pengguna.
2. Pencarian dokumen dengan menggunakan algoritma *Centroid Linkage Hierarchical Method* dengan pola analisa varian *Hill Climbing* dapat digunakan untuk mengelompokkan dokumen secara otomatis dengan jumlah *cluster* yang tepat.
3. Pola analisa varian dengan menggunakan metode *Hill Climbing* memerlukan waktu yang lebih cepat dalam melakukan analisa jumlah *cluster* jika dibandingkan dengan metode *valley tracing*. Hal ini disebabkan karena peng*cluster*an hasil dari *Hill Climbing* mendukung akses kecepatan penghitungan dokumen pada tiap *cluster*nya.
4. Pola analisa varian dengan menggunakan metode *Hill Climbing* sangat sesuai untuk pencarian dokumen dengan jumlah yang sangat besar dan kata kunci yang panjang. Hal ini berpotensi untuk implementasi program dalam skala yang lebih luas.

5.2. SARAN

Di dalam proyek akhir ini terdapat beberapa kelebihan dan kekurangan yang membutuhkan saran-saran untuk semakin mengembangkan proyek akhir ini sehingga bisa menjadi lebih sempurna. Adapun saran-saran yang diberikan adalah sebagai berikut:

1. Untuk pengembangan program ini, perlu untuk dicoba metode *Single Linkage Hierarchical Method* dalam melakukan proses *cluster*. Metode tersebut sangat cocok untuk dipakai pada kasus *shape independent clustering*, karena kemampuannya untuk membentuk *patern* tertentu dari *cluster*. Sehingga pembentukan *cluster* bisa lebih baik lagi.

2. Pada proses *text mining* diharapkan bisa untuk dilakukan pada bahasa lain seperti bahasa Arab yang memiliki struktur *morphological* yang lebih kompleks daripada bahasa Inggris. Sehingga kemampuan *text mining* akan semakin baik dan tentu penggunaanya akan lebih meluas.
3. Pada proses pencarian dan penyimpanan dokumen secara *online* masih terbatas pada *file* html yang diunduh dari Rss berita *online* di internet, diharapkan pengembangan berikutnya tidak hanya berupa *file* html saja, namun bisa bermacam *file* seperti: pdf, ppt, doc, dll. Sehingga *clustering*nya tidak hanya terbatas pada dokumennya saja namun diharapkan bisa juga *clustering* untuk *type* data dokumen sekaligus.
4. Label pada tiap kluster masih sebatas label hasil iterasi, misal *cluster* ke-1. Untuk pengembangan ke depan, pelabelan kluster bisa dilakukan secara otomatis sesuai dengan hasil dari tiap kluster yang mewakili kluster tersebut.

6. DAFTAR PUSTAKA

[1] A.R. Barakbah, K. Arai, *A New Algorithm For Optimization Of K-Means Clustering With Determining Maximum Distance Between Centroids*, In. IES 2006, Politeknik Elektronika Negeri Surabaya, ITS.

[2] Agusetia Usmaida, *WEB MINING UNTUK PENCARIAN BERDASARKAN KATA KUNCI DENGAN TEKNIK CLUSTERING*, Tugas Akhir Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya, Surabaya 2007.

[3] Porter MF (1980) *An algorithm for suffix stripping. Program*, 14: 130-137.

[4] A.R. barakbah, *Clustering*, In. Workshop Data Mining 2006, Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya,ITS.

[5] Hasniawati Helmy, *IMAGE CLUSTERING BERDASARKAN WARNA UNTUK IDENTIFIKASI BUAH DENGAN METODE VALLEY TRACING*, Tugas Akhir Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya, Surabaya 2007.

[6] A.R. Barakbah, K. Arai, *Identifying moving variance to make automatic clustering for normal data set*, In. Proc. IECI Japan Workshop 2004 (IJW 2004), Musashi Institute of Technology, Tokyo.